Transcriptional landscape estimation from tiling array data using a model of signal shift and drift

Pierre Nicolas1,∗, Aurélie Leduc1, Stéphane Robin2, Simon Rasmussen3, Hanne Jammer2 and Philippe Bessières1
1INRA, Mathématique Informatique et Génome UR1077, 78350 Jouy-en-Josas, 2AgroParisTech/INRA, Mathématiques et Informatique Appliquées UMR518, 16 rue Claude Bernard, 75005 Paris, France and 3Technical University of Denmark, Center for Biological Sequence analysis, Building 208, 2800 Lyngby, Denmark

Received on January 27, 2009; revised on May 11, 2009; accepted on June 19, 2009
Advance Access publication June 26, 2009
Associate Editor: Trey Ideker

ABSTRACT

Motivation: High-density oligonucleotide tiling array technology holds the promise of a better description of the complexity and the dynamics of transcriptional landscapes. In organisms such as bacteria and yeasts, transcription can be measured on a genome-wide scale with a resolution >25 bp. The statistical models currently used to handle these data remain however very simple, the most popular being the piecewise constant Gaussian model with a fixed number of breakpoints.

Results: This article describes a new methodology based on a hidden Markov model that embeds the segmentation of a continuous-valued signal in a probabilistic setting. For a computationally affordable cost, this framework (i) alleviates the difficulty of choosing a fixed number of breakpoints, and (ii) permits retrieving more information than a unique segmentation by giving access to the whole probability distribution of the transcription profile. Importantly, the model is also enriched and accounts for subtle effects such as signal ‘drift’ and covariates. Relevance of this framework is demonstrated on a Bacillus subtilis dataset.

Availability: A software is distributed under the GPL.

Contact: pierre.nicolas@jouy.inra.fr

Supplementary information: Supplementary data is available at Bioinformatics online.

1 INTRODUCTION

High-density oligonucleotide tiling arrays carry tightly spaced probes that provide uniform covering of the genomic sequence. By hybridization with RNA samples (cDNA), they have been used to query the transcriptional activity of the whole genome in an array of model organisms (Bertone et al., 2004; Biemann et al., 2006; He et al., 2007; Stolc et al., 2005). The approach is particularly attractive for organisms with small-sized genome such as bacteria and yeasts where a resolution >25 bp is more easily achieved (David et al., 2006; Rasmussen et al., submitted for publication). The generalization of the use of such arrays should provide unbiased and high-quality pictures of the complexity and the dynamics of transcriptional landscapes (Xu et al., 2009). The great promise of these data justifies the improvement of the currently available statistical methods dedicated to their analysis.

From the methodological standpoint, the problem is naturally stated in terms of finding segments where the hybridization signal is relatively constant, delimited by breakpoints that are expected to correspond to biological features such as transcript start and stop sites or splicing sites. A variety of tools including local non-parametric smoothing (Royce et al., 2007; Wang et al., 2009) and iterative hypothesis testing (Olshen et al., 2004) have been proposed to answer this question. Probably the most popular and best mathematically grounded methodology consists of seeking the piecewise constant model with Gaussian noise that best fits the signal (Huber et al., 2006; Picard et al., 2005). Namely, for a fixed number of segments S, fitting the model consists of finding the combination of breakpoints 1 < t1 ≤ ··· ≤ tS−1 ≤ n that minimizes the sum of squared residuals:

\[ G(t_1, \ldots, t_{S-1}) = \sum_{k=1}^{n} \sum_{l=t_{k}+1}^{t_{k+1}} (x_l - \bar{x}_k)^2, \]

where \( x_k \) is the signal at position \( k \), \( \bar{x}_k \) is the average signal level in segment \( s \) (i.e. between \( t_{k-1} \) and \( t_{k} \)), \( t_0 = 1 \) and \( t_{S} = n+1 \).

In full generality, minimizing the sum of squared residuals in Equation (1) can be achieved by Dynamic Programming and requires time \( O(n^2S) \). Huber et al. (2006) fixed an upper bound \( l \) on the maximum length of each segment to reduce the time complexity to \( O(nlS) \) with \( l < n \). The problem of choosing the correct number of segments \( S \) was more specifically examined by Picard et al. (2005), but visual assessment and use of prior belief have also been advocated (Huber et al., 2006) and have been useful in practice (David et al., 2006, Rasmussen et al., submitted for publication).

The simplicity of this approach is appealing but hides a number of difficulties, the most important being the choice of the number of segments. In principle, this issue can be tackled by embedding the segmentation model in a probabilistic setting that includes not only the noise but also the evolution of the signal. This idea stimulated the development of hidden Markov models (HMMs) (Fridlyand et al., 2004; Marioni et al., 2006; Stjernqvist et al., 2007) for the analysis of comparative genomic hybridization data. For transcriptomic data, a different approach consists of training HMMs to distinguish between transcribed and non-transcribed regions (Da et al., 2006; Munch et al., 2006).
et al., 2006). When the quality of the data is good enough it is both more natural and more ambitious to try to recover the ‘denoised’ transcription signal instead of directly summarizing the data via a classification algorithm. Transcript level is, however, a continuous quantity and none of the available models is satisfactory for a continuous-valued underlying signal. An HMM that achieves this aim at a computationally affordable cost is described in the present article. The proposed model does also extend the piecewise constant model in two directions. First, it integrates the influence of covariates that serve to account for differential affinity between probes. This allows to achieve segmentation and within-array normalization in one step. Second, the proposed model relaxes the assumption of strictly constant transcript levels between abrupt ‘shifts’ by also allowing progressive ‘drift’ of the signal. Inference based on this model is examined and discussed.

2 METHODS

2.1 Experimental data

The main example dataset used here comes from pilot experiments conducted on Bacillus subtilis within the European Consortium BioSyBio (S.Rasmussen et al., submitted for publication). This array consists of 383,149 probes starting every 22 nt on each strand of the B. subtilis genome (GenBank: AL010916). Probe lengths range between 45 nt and 65 nt and were adjusted to reduce melting temperature (TM) variations (isothermal design). Production of the tiling arrays, synthesis of labeled cDNA from (GenBank: AL009126). Probe lengths range between 45 nt and 65 nt and were carried out by Nimblegen. Antisense artifacts were controlled by using design). Production of the tiling arrays, synthesis of labeled cDNA from

2.2 Shift and drift in an HMM framework

Like in previous approaches (Huber et al., 2006; Ohlen et al., 2004; Picard et al., 2005), the log2 of the observed intensity \( y_i \) is modeled as the sum of an unobservable signal \( u_i \) that is the focus of interest plus a Gaussian noise with SD \( \sigma \). This general model can be written as:

\[
x_i | u_i \sim \mathcal{N}(u_i, \sigma^2).
\]

However, \( u_i \) is not seen in our model as a parameter but is itself a random variable. Correlation between probes that are adjacent on the chromosome is accounted for by a Markov transition kernel \( \pi(u_{i+1}, u_i) \) and \( x_i, u_{i+1} | u_i \)

\[
\pi(u_{i+1}, u_i) = 
\begin{cases} 
\alpha \left[ \left( \frac{u_{i+1} - u_i}{\delta} \right)^2 \right]^{\frac{1}{2\alpha}} & \text{if } |u_{i+1} - u_i| \leq \delta \\
0 & \text{otherwise}
\end{cases}
\]

where \( \alpha \) is the kernel density estimate computed on \( x \) with a Gaussian kernel and Scott’s bandwidth (Scott, 1992). The possibility of small drift, either upward or downward, is accounted for by \( u_a \) and \( u_d \). Drift amplitudes are modeled by two geometric distributions of parameters \( \lambda_a \) and \( \lambda_d \) and average amplitudes write \( h = h(1 - \lambda) \).

It can be verified that as \( h \to 0 \) and \( h(1 - \lambda) \to 0 \) the transition kernel of the discrete-valued Markov chain of Equation (3) converges in distribution toward the transition kernel of a continuous-valued Markov chain. In its continuous form, the kernel writes as a mixture of a point mass at \( u_i \) of weight \( a \), a continuous-valued distribution of density \( y \) and weight \( a \), and two shifted exponential distributions of rates \( y_\beta \) and \( y_\gamma \) and weights \( a_u \) and \( a_v \). With an appropriately high \( K \) it should thus be possible to approach, using the discrete-valued model of Equation (3), the results that one would obtain with the continuous-valued model.

The Supplementary Material available online gives a detailed presentation of the equations that allow OncoK implementations of the HMM classical algorithms, namely:

(1) Log-likelihood computation \( \mathbb{P}(x_t, u_{-t}) \)
(2) Forward-backward algorithm (computation of \( \mathbb{P}(x_t | u_{-t}) \) for each \( t \))
(3) Viterbi algorithm (finding the trajectory \( u_{-t} | x_{-t} \) that maximizes \( \mathbb{P}(x_t | x_{-t}) \)).

These algorithms are implemented in our software. All the parameters are estimated in the maximum likelihood (ML) framework with the EM algorithm, an iterative algorithm that alternates an E-step (forward-backward algorithm) and a M-step (parameter update). The output provides a detailed report on the ‘denoised’ signal based on the results of the Viterbi and forward-backward algorithms.

2.3 gDNA signal as a covariate

gDNA hybridization data were used in a preprocessing step by Huber et al. (2006) for the purpose of between-probe signal normalization and outlier analysis.
trimming. The model proposed here accounts for these effects by modeling the gDNA hybridization intensities as a covariate.

The probability distribution for the observed variable $x_t$ given the underlying signal $u_t$ and the gDNA residuals $r_t$ writes as a mixture model

$$x_t | u_t, r_t \sim (1 - \varepsilon(r_t))N(u_t + \rho(r_t)\sigma^2, \sigma^2) + \varepsilon(r_t)\delta(U_{\text{max}}, U_{\text{min}}),$$

where $\varepsilon(r_t)$ corresponds to the probability of outliers, $U(U_{\text{max}}, U_{\text{min}})$ is the uniform distribution that models outlier data and $N(u_t + \rho(r_t)\sigma^2, \sigma^2)$ is the Gaussian distribution modeling non-outlier data. This model is markedly richer than Equation (2). Notice (i) the non-constant proportionality factor $\rho(r_t)$ applied to $r_t$, (ii) the non-constant standard error $\sigma(u_t)$ of the Gaussian distribution; and (iii) the probability of outliers $\varepsilon$ that depends on $r_t$. More precisely, $\rho$ and $\sigma$ are modeled as piecewise constant function of $u_t$ with eight intervals, and $\varepsilon$ is a two-parameter logistic function of the absolute value of $u_t$, $\varepsilon(u_t) = 1/(1 + e^{-\alpha - \beta u_t})$. All the parameters are simultaneously estimated with the EM algorithm (see Supplementary Material).

Finally, left and right censoring are incorporated in the model to account for the experimental limitations that preclude exact measurements of extremely high and extremely low intensities. In practice, the lower and upper 5% of the original range of variation of the intensity $x_t$ are considered as censored.

### 3 RESULTS AND DISCUSSION

#### 3.1 Selecting the appropriate level of discretization

The model was designed with the explicit aim of modeling a continuous-valued underlying signal. In other words, discretization of the hidden state space is seen only as a necessary technicality and the step $h = 1/K$ should ideally be sufficiently small to have no impact on the results. Intuitively, the smaller the SD of the noise $\sigma$, the smaller the step $h$ should be. The results obtained on the

![Fig. 1. Influence of the number of hidden states. (A) Log-likelihood (in natural log) as a function of the number of hidden states, $K$. (B) Estimated average variance of the noise $\theta$ as a function of $K$ (plain line). The discretization step $h \propto 1/(K - 1)$ is also shown (dotted line).](image-url)

B.subtilis dataset and presented in Figure 1 confirm this intuition and thereby provide some form of validation for the model. Figure 1 shows that increasing $K$ (and thus decreasing $h$) actually increases the model adequation to the data as measured by the log-likelihood after ML estimation. Beyond a certain value of $K$ the impact of this change becomes, however, almost unnoticeable. Figure 1 also reports the parallel evolution of $h$ and $\sigma$. According to this plot, having $h$ around 0.5 seems more than sufficient. Indeed, with such a value of $h$, the 95% confidence interval ($\text{CI}$) of the distribution of the noise is about eight times as large as the discretization interval $h$. $K$ was set to 100 for this particular dataset.

This choice of $K = 100$ corresponds to an acceptable running time for the algorithm. Our setting throughout this study consisted to explore 10 random starting points for the EM algorithm. Here, it resulted in a total of 885 iterations taking 5 h 6 min on an Intel(R) Xeon(TM) CPU 3.40 GHz CPU, less than the 5 h 36 min needed for the segmentation algorithm of Huber et al. (2006) with maximum segment length $l = 1000$ (22 000 bp) and segment number on each strand $S = 1500$.

#### 3.2 Importance of modeling drift and covariates

Parameter estimates in model-based analyses are an invaluable source of information to understand both the behavior of the model and the data. The model contains a total of 23 parameters. Figure 2 is intended to provide an overview of their ML estimates on the B.subtilis data. The first row of Table 1 gives numerical values for the different parameter estimates. More precisely, the value $\alpha$ of the parameter $\alpha$ is $0.095$. The sharp peak reflects the high value of $\alpha$; it is estimated that the underlying signal remains

![Fig. 2. Parameter estimates. (A) Transition matrix $\pi(u_t, u_{t+1})$. One row is represented. (B) SD of the noise $\sigma$ as a function of the signal level $u_t$. (C) Outlier probability $\varepsilon$ as a function of the magnitude of the gDNA residuals $r_t$ (plain line) and complementary cumulative distribution function of the gDNA residuals (dotted line). (D) Proportionality factor $\rho$ applied to $r_t$ as a function of the signal level $u_t$.](image-url)
Table 1. Model comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>(d^a)</th>
<th>(\beta^1)</th>
<th>(a_1)</th>
<th>(a_2)</th>
<th>(\rho)</th>
<th>(\hat{\alpha})</th>
<th>CV-LL^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>23</td>
<td>y</td>
<td>0.012</td>
<td>0.050</td>
<td>0.078</td>
<td>0.66</td>
<td>0.30</td>
</tr>
<tr>
<td>M2</td>
<td>21</td>
<td>y</td>
<td>0.016</td>
<td>0.046</td>
<td>0.074</td>
<td>0.66</td>
<td>0.31</td>
</tr>
<tr>
<td>M3</td>
<td>21</td>
<td>y</td>
<td>0.029</td>
<td>0</td>
<td>0.039</td>
<td>0.64</td>
<td>0.33</td>
</tr>
<tr>
<td>M4</td>
<td>21</td>
<td>y</td>
<td>0.040</td>
<td>0.014</td>
<td>0</td>
<td>0.64</td>
<td>0.34</td>
</tr>
<tr>
<td>M5</td>
<td>19</td>
<td>y</td>
<td>0.046</td>
<td>0</td>
<td>0</td>
<td>0.63</td>
<td>0.34</td>
</tr>
<tr>
<td>M6</td>
<td>15</td>
<td>y</td>
<td>0.012</td>
<td>0.156</td>
<td>0.197</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>M7</td>
<td>15</td>
<td>y</td>
<td>0.014</td>
<td>0.036</td>
<td>0.053</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>M8</td>
<td>9</td>
<td>y</td>
<td>0.046</td>
<td>0</td>
<td>0</td>
<td>0.50</td>
<td>0.50</td>
</tr>
</tbody>
</table>

^aModel dimension as measured by the number of free parameters.
^1y if the model accounts for outliers, ‘n’ otherwise.
^2CV-LL: cross-validated log-likelihood.

The narrow shoulders on both sides of the peak correspond to the upward and downward shift moves and reflect the value of the parameters \((a_1, \lambda_0)\) and \((a_2, \lambda_d)\), respectively. Close inspection reveals a small asymmetry, with upward moves being less frequent than downward moves (5.0% versus 7.8%). The small estimated proportion of abrupt shift moves between adjacent probes is almost invisible at this scale (1.2%).

As expected, the probability of outliers is estimated to increase with the magnitude of the residuals of the gDNA signal. The two-parameter logistic curve that models this relationship is shown in Figure 2C. Remarkably, the probability of outliers is just slightly below unity, its sharp decrease for low values of the signal. This behavior is a relatively flat function in the middle of the spectrum is just slightly below unity, the value that we expect in an idealized situation [see the rationale behind the preprocessing step in Huber et al. (2006)].

As a whole, these results emphasize the importance of two specificities of our model: the modeling of drift moves as a complement to shift moves and the non-constant \(\rho\) that provides a simple adaptive method to account for the variation of affinity between probes.

To better understand the behavior of the model and the characteristics of the data, we carried out a comparative analysis of eight models. For the purpose of robust assessment of model fitness with respect to the B. subtilis dataset each model was fitted two times, once on each strand of the chromosome, and the likelihood was each time computed on the other strand. The sum of both log-likelihood terms is reported as the cross-validated log-likelihood in Table 1. Parameter values in Table 1 were estimated on the full dataset.

Sorted by decreasing value of adequacy with the data, the models ranged from M1, the full 23-parameter model, to M8, a nine-parameter model that does not account for drifts, outliers nor covariates. Not accounting for outliers has only a small impact on the overall model fitness (M2 versus M1), but the probability of shift moves is increased by 30% in this simpler model. This can have a non-negligible impact in practice given that these particular shift moves are indeed likely to be spurious. Not modeling drifts has a much more pronounced impact (M5 versus M1). Fitness is 6.5% better for M1 than for M5 and the estimated proportion of shift moves is about four times lower in M1 (1.2% versus 4.6%), suggesting that a substantial fraction of the drift moves in M1 are interpreted as shift moves in M5. A closer examination underscores the importance of downward drift as compared with upward drift. Not accounting for downward drift has 74% more effect on the overall fitness that not accounting for upward drift (M5 and M4 versus M1). More spectacularly, if a single drift direction is allowed, modeling downward drift improves the model ~4.5 times more than modeling only upward drift (M3 versus M5 and M4 versus M5). Setting \(\rho\) to either 1 or 0 were both found to result in a dramatic drop in fitness but with different specific effects. Setting \(\rho\) to 1 in M6 results in estimation of high drift compared with original model, whereas setting \(\rho\) to 0 in M7 results in estimation of high noise.

3.3 Estimation of transcriptional landscape: illustration on B. subtilis data

The ultimate goal of the use of the model is to infer the underlying signal supposed to reflect the actual transcriptional landscape. The adoption of a probabilistic setting for the trajectory of the underlying signal allows for a considerably richer signal reconstruction than just ‘optimal’ trajectory reconstruction. Figure 3 gives an illustration of these possibilities by superimposing a number of results obtained with the model on a 10 000 bp region of the B. subtilis chromosome. Results include: (i) the prediction interval for the value of the signal \(u_t\) at each chromosome position; (ii) a point prediction for the signal value by the conditional mean of \(u_t\) (the best predictor in terms of quadratic error); (iii) the inferred position of the experimental point after correction for differential probe affinity [computed as \(u_t - \hat{\mu}(u_t)|\xi]\); (iv) the exact position of each type of move in the best trajectory given by the Viterbi path (abrupt shift, upward drift and downward drift); and (v) the probability of having each type of move at each position. All these values can be read directly from the output of our software.

The biological pertinence of the distinction between shifts and drifts seems remarkable in Figure 3. Inferred shifts are found mostly in intergenic regions that a priori correspond to possible positions for transcriptional promoters and terminators. The position of each move (2893 shifts and 13 460 drifts) was compared with sequence predictions for two biological features: Rho-independent (intrinsic) terminators predicted with the algorithm of d’Aubenton-Carafa et al. (1990); promoters dependent on Sigma-A predicted using an HMM whose structure was chosen according to the results of Nicolas et al. (2006). To fulfill the needs of an unbiased analysis, both categories of predictions were made without prior on the position of the genes and confidence cutoffs were set relatively low to increase sensitivity (a total 4164 Sigma-A predictions and 3492 terminator predictions are considered).

The results presented in Figure 4 confirm the practical relevance of the distinction between shift and drift moves. For upward moves, it shows the difference between shift and drift with respect to the distance between the breakpoint and the nearest promoter prediction.
Probabilistic modeling of tiling array expression data

Fig. 3. Transcriptional landscape inference. Analysis of the signal on the (+)-strand of a 10 000 bp segment of the *B. subtilis* chromosome. Upper part: open circles show the original signal. Closed gray circles represent the signal after ‘correction’ with the gDNA covariate. The thick black line shows the expectation of the transcript level as computed with the HMM. Thin black lines correspond to the 95% CI. Middle part: horizontal arrows indicate GenBank CDSs. Lower part: shift moves along the most likely trajectory are shown as squares. Upward and downward drift moves are indicated by point-up and point-down triangles, respectively. Move probabilities are represented as gray lines.

Drift might partly reflect local variations of labeled cDNA that result from technical artifacts such as random priming bias. Drift could also reflect biological differences in the amount of mRNA. In particular, Figures 4 and S1 leave no doubt that a fraction of the drifts correspond to promoters or terminators whose activity is too weak to be detected as shifts in this biological condition. A preliminary exploration of the patterns of drift is reported in the Supplementary Material. Figure S2 shows that downward drift is most pronounced after upward shifts and before downward shifts, near the 5’ and 3’ ends of transcriptionally active regions. An excess of upward drift is found before upward shifts, at the 3’ end of regions with low-transcriptional activity. Random priming artifacts could most easily be invoked to explain downward drift at the 3’ end of transcriptionally active regions. Downward drift may also, for instance, be partly caused by molecules whose synthesis is still incomplete. Here, no single explanation could apparently account the patterns of upward and downward drift. Instead, drift is observed in a variety of chromosomal and transcriptional contexts that the landscape snapshots presented in Figures S3, S4 and S5 intend to illustrate. As an example, some spectacular cases of downward drift are found for transcription units apparently lacking a clear terminator. The intensity of the resulting downstream antisense transcription drifts downward progressively. In Figure S3, a pattern reminiscent of the bidirectional transcriptional activity recently described in *S. cerevisiae* (Xu et al., 2009) can also be observed.

3.4 Benchmark comparisons

In addition to allow insightful reconstructions of the transcriptional landscape, good algorithms should identify breakpoints that match, as closely as possible, the position of the promoters and terminators. To compare different sets of breakpoints, promoter and terminator
predictions were used as a proxy for the true (unknown) reference. Results are shown in Figure 5.

The results obtained with the HMMs, \( A1, A5, A15 \) and \( M8 \), give another confirmation of the biological pertinence of the distinction between shift moves and drift moves in \( M1 \). It also revealed the deep impact of the correction for variation of affinity between probes using covariates, not implemented in \( M8 \). The misbehavior of \( M8 \) translates paradoxically in an apparent success at detecting terminators. This most likely does not reflect the transcription signal itself, but rather the low probe affinity due to the stem–loop secondary structure distinctive of the rho-independent terminators.

For the comparison of the new HMM segmentation method and the piecewise constant regression implemented in the algorithm of Huber et al. (2006), the later was run on the data after correction for difference of affinity between probes (as shown in Fig. 3) with maximum segment length \( l = 22000 \) bp and number of segments on each strand \( S = 1000 \) and 3000. Results clearly demonstrate the benefit of the new HMM framework. For \( S = 1500 \), the number of breakpoints matching promoter and terminator predictions were, respectively, 8.9% and 25% higher for the HMM.

3.5 Results on \( S.cerevisiae \) data

Examination of the segmentation produced by piecewise constant regression on Watson (+)-strand of \( S.cerevisiae \) yeast chromosome 1 leads to the choice of 152 (average segment size 1500bp) as a sensible number of breakpoints (Huber et al., 2006). A question was thus whether the automatic procedure presented here will identify a similar number of shift moves. The model was fitted on the mRNA and gDNA data of the 57 616 probes representing both strands of the chromosome 1.

The Viterbi path of our HMM on the (+)-strand contained 125 shift moves and 373 drift moves with a median distance of 60bp between each of the 152 breakpoints of Huber et al. (2006) and the closest of the 125 shift moves. On this dataset, modeling drift can thus be useful to single out the most abrupt changes in the signal intensity.

Interestingly, further comparisons of models with and without drift indicated that drift improve the model fitness by only 1.4% on the \( S.cerevisiae \) data, much less than the 6.5% found on \( B.subtilis \) data. Biology and array technology are two sources of possible differences between \( S.cerevisiae \) and \( B.subtilis \) datasets. Our model of drift seems more relevant for prokaryotic data obtained using long isothermal probes.

4 CONCLUSIONS

This article describes a new methodology based on an HMM that embeds the segmentation of a continuous-valued signal in a probabilistic setting. For a computationally affordable cost, this framework alleviates the difficulty of choosing a fixed number of breakpoints and permits retrieving more information than a unique segmentation. Probabilistic modeling makes it straightforward to compute confidence measures on the estimated transcriptional landscape. This information should prove particularly useful to pinpoint the differences in large collections of arrays. Extension of the model could also be imagined to tackle the problem of the joint segmentation of datasets where transcript boundaries and expression level differ.

By accounting for gDNA hybridization data as a covariate, the model automatically corrects the data for the variation of affinity between probes. David et al. (2006) proposed for this purpose a preprocessing step to be carried out on the raw data, before log-transformation, and producing a significant fraction of negative values. The data could thus no longer be simply log-transformed and more complicated variance stabilization transformation, requiring multiple arrays, was used (Huber et al., 2002). In comparison, the normalization carried out by the model needs only one array and it alters only minimally the overall distribution of the log of the original data.

The model is also enriched and accounts for subtle effects such as signal ‘drift’ and covariates. Interestingly, our results unambiguously document the existence of a drifts in the \( B.subtilis \) dataset. The interest of this observation is 2-fold. First, drift have not been accounted in the previous models and this may partially explain the misbehavior of \( M8 \) translates paradoxically in an apparent success at detecting terminators. This most likely does not reflect the transcription signal itself, but rather the low probe affinity due to the stem–loop secondary structure distinctive of the rho-independent terminators.

Second, the causes and the patterns of drift deserve to be investigated if we want to make the best use of tiling array expression data. The software is distributed under the GNU Public License http://genome.jouy.inra.fr/~pnico/ausubetc/.

Acknowledgements

We thank Etienne Dervyn, Philippe Noirot and Franck Picard for constructive comments on the content of the manuscript.

Funding: BaSysBio project, European Commission research grant (LSHG-CT2006-037469). Conflict of Interest: none declared.

REFERENCES

Probabilistic modeling of tiling array expression data


